Safe and efficient off-policy reinforcement learning

Review by Alexis Sair & Antoine Hoorelbeke

Abstract

In this paper, we review *Safe and efficient off-policy reinforcement learning* by Rémi Munos, Thomas Stepleton, Anna Harutyunyan and Marc G. Bellemare.
The article is interested in a new off-policy reinforcement learning algorithm called Retrace(λ) which improves other well-known off-policy algorithms. Results on the contraction properties of this algorithm for both policy evaluation and control are proved. The authors also show that the algorithm can learn from sample trajectories, in an online setting.

In our review, we adopt both a theoretical and a practical point of view : we explain and provide additional arguments for the three theorems and implement the Retrace(λ) algorithm in the frozen-lake environment.

1 Notation

We use the same notations as in the article. We are in the context of a Markov Decision Process $(\mathcal{X}, \mathcal{A}, \gamma, P, r)$ where $\mathcal{X}$ stands for the finite state space, $\mathcal{A}$ is the action space (considered finite), γ is the discount factor in $[0, 1)$, P is the transition function, mapping each couple $(x, a) \in \mathcal{X} \times \mathcal{A}$ to a distribution over $\mathcal{A}$ and $r : \mathcal{X} \times \mathcal{A} \to [-R_{\text{MAX}}, R_{\text{MAX}}]$ is the reward function.

Like in the course, for a policy π, Q^π will denote the expected discounted reward associated with following π from a given state-action pair. More generally we will call Q-function any function from $\mathcal{X} \times \mathcal{A}$ to R. We introduce for a Q-function Q the P^π operator :

$$\left(P^\pi Q\right)(x, a) := \sum_{x' \in \mathcal{X}} \sum_{a' \in \mathcal{A}} P\left(x'|x, a\right) \pi\left(a'|x'\right) Q\left(x', a'\right)$$

and

$$Q^\pi := \sum_{t \geq 0} \gamma^t \left(P^\pi\right)^t r \tag{1}$$

Let's explain that relation quickly. If we start from an initial state-action (x, a) at time 0, the expected discounted reward at time 1 is

$$\gamma \sum_{x' \in \mathcal{X}} \sum_{a' \in \mathcal{A}} P\left(x'|x, a\right) \pi\left(a'|x'\right) r\left(x', a'\right)$$

that is $\gamma(P^\pi r)(x, a)$. By induction the expected discounted reward at time t is $\gamma^t (P^\pi)^t r$. That gives the expression of Q^π with the operator P^π.

The *Bellman operator* $\mathcal{T}^\pi$ for a policy π is defined as $\mathcal{T}^\pi Q := r + \gamma P^\pi Q$. Let's check that it admits Q^π for a fixed point. From (1) we have that

$$\begin{aligned}
\mathcal{T}^\pi Q^\pi &= r + \gamma P^\pi Q^\pi \\
&= r + \sum_{t \geq 0} \gamma^{t+1} P^{\pi(t+1)} r \\
&= r + Q^\pi - r \\
&= Q^\pi
\end{aligned}$$

and we check easily (since $\gamma \|P^\pi\| < 1$ makes $I - \gamma P^\pi$ invertible) that $(I - \gamma P^\pi) Q^\pi = r$ that is $Q^\pi = (I - \gamma P^\pi)^{-1} r$
We can then introduce the *Bellman optimal operator*

$$\mathcal{T} Q := r + \gamma \max_\pi P^\pi Q$$

whose fixed point is the optimal state-value function Q^* we look foor. We also introduce the λ return based operators

$$\mathcal{T}_\lambda^\pi Q := (1 - \lambda) \sum_{n \geq 0} \lambda^n \left[\left(\mathcal{T}^\pi\right)^n Q\right] = Q + \left(I - \lambda \gamma P^\pi\right)^{-1} \left(\mathcal{T}^\pi Q - Q\right)$$

which obviously admit Q^π as fixed point since it is a fixed point of $\mathcal{T}^\pi$. In the following we will denote as μ the behavior policy and $\mathcal{F}_t$ the tribu up to time t.

2 Off-policy algorithms

We introduce the $\mathcal{R}$ operator such as

$$\mathcal{R}Q(x,a) := Q(x,a) + E_\mu\left[\sum_{t\geq 0}\gamma^t\left(\prod_{s=1}^{t} c_s\right)\left(r_t + \gamma E_\pi Q\left(x_{t+1},\cdot\right) - Q\left(x_t,a_t\right)\right)\right] \tag{2}$$

If we look closer at the operator R, we see that it is the usual form of algorithms in the reinforcement setting : we apply the identity and add a weighted error. Here the weighted error is the expectation with respect to the behavior policy μ of the discounted error at time t : $r_t + \gamma E_\pi Q\left(x_{t+1},\cdot\right) - Q(x_t,a_t)$. The product of the c_s adds a degree of freedom on which we want to play to improve the algorithm. We also notice that Q^π is a fixed point of R because $E_{x_{t+1}\sim P(\cdot|x_t,a_t)}\left[r_t + \gamma E_\pi Q^\pi\left(x_{t+1},\cdot\right) - Q^\pi\left(x_t,a_t\right)\right] = \mathcal{T}^\pi Q^\pi - Q^\pi = 0$ since we saw that Q^π was a fixed point of $\mathcal{T}^\pi$.

Importance sampling (IS): $c_s = \frac{\pi(a_s|x_s)}{\mu(a_s|x_s)}$. Let's show that this choice leads to $\mathcal{R}Q = Q^\pi\ \forall Q$. To show this, we use **Lemma 1**, that we will demonstrate later.

Lemma 1 The difference between $\mathcal{R}Q$ and its fixed point Q^π is

$$\mathcal{R}Q(x,a) - Q^\pi(x,a) = E_\mu\left[\sum_{t\geq 1}\gamma^t\left(\prod_{i=1}^{t-1} c_i\right)\left(\left[E_\pi\left[\left(Q-Q^\pi\right)\left(x_t,\cdot\right)\right] - c_t\left(Q-Q^\pi\right)\left(x_t,a_t\right)\right]\right)\right]$$

Lemma 0 For IS, $\mathcal{R}Q = Q^\pi\ \forall Q$

Proof For importance sampling we have $\forall s$, $c_s = \frac{\pi(a_s|x_s)}{\mu(a_s|x_s)}$. As a consequence, with $C_t := \prod_{i=1}^{t} c_i$, we have :

$$\begin{aligned} E_\mu\left(C_t\left(E_\pi\left[\left(Q-Q^\pi\right)\left(x_t,\cdot\right)\right] - c_t\left(Q-Q^\pi\right)\right)\left(x_t,a_t\right)\right) &= E_{x_{1:t},a_{1:t}}\left[C_{t-1}\left[E_\pi\Delta Q\left(x_t,\cdot\right) - c_t\Delta Q\left(x_t,a_t\right)\right]\right] \\ &= E_{x_{1:t},a_{1:t-1}}\left[C_{t-1}\left(E_\pi\Delta Q(x_t,\cdot) - E_{a_t}(c_t\Delta Q(x_t,a_t))\right)\right] \end{aligned}$$

by using **Lemma 1**

$$\begin{aligned} E_\pi\Delta Q(x_t,\cdot) - E_{a_t}(c_t\Delta Q(x_t,a_t)) &= \sum_{b\in\mathcal{A}}\Delta Q(x_t,b)\pi(b|x_t) - \frac{\pi(b|x_t)}{\mu(x_t,b)}\mu(x_t,b)\Delta Q(x_t,b) \\ &= 0 \end{aligned}$$

since $a_t \sim \mu(\cdot, x_t)$.

So we have, in the context of Importance Sampling that $\mathcal{R}Q = Q^\pi \ \forall Q$.
The main problem of Importance Sampling is the variance of the estimates since $\frac{\pi}{\mu}$ may take big values.

Off-policy $\mathbf{Q}^\pi(\lambda)$ and $\mathbf{Q}^*(\lambda)$: $c_s = \lambda$ It can be shown that the variance is reduced and that when μ and π are close, $\mathcal{R}$ is a contraction mapping around Q^π and Q^* under certain conditions. That allows convergence properties.

Tree-backup, $\mathbf{TB}(\lambda) : c_s = \lambda\pi\left(a_s|x_s\right)$

Retrace$(\lambda) : c_s = \lambda \min\left(1, \frac{\pi(a_s|x_s)}{\mu(a_s|x_s)}\right)$ This is the values of the traces we will be interested in the article.

3 Analysis of Retrace(λ)

3.1 Policy evaluation

Given a policy π, we want to evaluate the state-value function associated Q^π by using the Retrace(λ) algorithm. In fact, we want to have convergence properties of the algorithm defined by induction with $\mathcal{R}$, so we show contraction properties around Q^π. Let's first show **Lemma 1**, we stated earlier.
Proof (Lemma 1). We have seen that Q^π is a fixed point of $\mathcal{R}$ s.t $\mathcal{R}Q^\pi = Q^\pi$. We first notice that

$$Q(x,a) + E_\mu\left(\sum_{t\geq 0}\gamma^t \prod_{s=1}^{t}\left(r_t + \gamma E_\pi Q(x_{t+1} - Q(x_t,a_t)\right)\right) =$$

$$Q(x,a)+E_\mu\left(\sum_{t\geq 0}\gamma^t\left(\prod_{s=1}^{t}c_s\right)\left(r_t+\gamma E_\pi Q(x_{t+1},\cdot)\right)\right)-E_\mu\left(\sum_{t\geq 0}\gamma^t\left(\prod_{s=1}^{t}c_s\right)Q(x_t,a_t)\right) =$$

$$E_\mu\left(\sum_{t\geq 0}\gamma^t\left(\prod_{s=1}^{t}c_s\right)\left(r_t+\gamma E_\pi Q(x_{t+1},\cdot)\right)\right)-E_\mu\left(\sum_{t\geq 1}\gamma^t\left(\prod_{s=1}^{t}c_s\right)Q(x_t,a_t)\right) =$$

$$E_\mu\left(\sum_{t\geq 0}\gamma^t\left(\prod_{s=1}^{t}c_s\right)\left(r_t+\gamma E_\pi Q(x_{t+1},\cdot)\right)\right)-E_\mu\left(\sum_{t\geq 0}\gamma^{t+1}\left(\prod_{s=1}^{t+1}c_s\right)Q(x_{t+1},a_{t+1})\right) =$$

$$E_\mu\left(\sum_{t\geq 0}\gamma^t\left(\prod_{s=1}^{t}c_s\right)\left(r_t+\gamma E_\pi Q(x_{t+1},\cdot)-\gamma c_{t+1}Q(x_{t+1},a_{t+1})\right)\right)$$

We may then write :

$$\mathcal{R}Q(x,a) = \sum_{t\geq 0} \gamma^t E_\mu \left[\left(\prod_{s=1}^{t} c_s \right) \left(r_t + \gamma \left[E_\pi Q\left(x_{t+1}, \cdot\right) - c_{t+1} Q\left(x_{t+1}, a_{t+1}\right) \right) \right] \right]$$

and then, using that Q^π is a fixed point of $\mathcal{R}$:

$$\begin{aligned}
\mathcal{R}Q(x,a) - Q^\pi(x,a) &= \sum_{t\geq 0} \gamma^t E_\mu \left[\left(\prod_{s=1}^{t} c_s \right) \left(\gamma \left[E_\pi \Delta Q\left(x_{t+1}, \cdot\right) - c_{t+1} \Delta Q\left(x_{t+1}, a_{t+1}\right) \right] \right) \right] \\
&= \sum_{t\geq 1} \gamma^t E_\mu \left[\left(\prod_{s=1}^{t-1} c_s \right) \left(\left[E_\pi \Delta Q\left(x_t, \cdot\right) - c_t \Delta Q\left(x_t, a_t\right) \right] \right) \right]
\end{aligned}$$

which is **Lemma 1**.

We use it to prove our first theorem we states that $\mathcal{R}$ is a contraction mapping around Q^π.

Theorem 1. The operator $\mathcal{R}$ defined by (3) has a unique fixed point Q^π. Furthermore, if for each $a_s \in \mathcal{A}$ and each history $\mathcal{F}_s$ we have $c_s = c_s\left(a_s, \mathcal{F}_s\right) \in \left[0, \frac{\pi(a_s|x_s)}{\mu(a_s|x_s)}\right]$, then for any Q-function Q

$$\|\mathcal{R}Q - Q^\pi\| \leq \gamma \|Q - Q^\pi\|$$

Proof. Using **Lemma 1** we have :

$$\begin{aligned}
\mathcal{R}Q(x,a) - Q^\pi(x,a) &= \sum_{t\geq 1} \gamma^t E_{a_{1:t}, x_{1:t}} \left[\left(\prod_{i=1}^{t-1} c_i \right) \left[E_\pi \Delta Q\left(x_t, \cdot\right) - c_t \Delta Q\left(x_t, a_t\right) \right] \right] \\
&= \sum_{t\geq 1} \gamma^t E_{a_{1:t-1}, x_{1:t}} \left[\left(\prod_{i=1}^{t-1} c_i \right) \left[E_\pi \Delta Q\left(x_t, \cdot\right) - E_{a_t}\left[c_t\left(a_t, \mathcal{F}_t\right) \Delta Q\left(x_t, a_t\right) | \mathcal{F}_t \right] \right] \right] \\
&\quad \text{where we use that } \forall \text{ F } \subset \mathcal{G}, \mathbf{E}(\mathbf{E}(X|\mathcal{G})|\mathcal{F}) = \mathbf{E}(X|\mathcal{F}) \\
&= \sum_{t>1} \gamma^t E_{a_{1:t-1}, x_{1,t}} \left[\left(\prod_{i=1}^{t-1} c_i \right) \sum_b \left(\pi(b|x_t) - \mu(b|x_t) c_t(b, \mathcal{F}_t) \right) \Delta Q\left(x_t, b\right) \right]
\end{aligned}$$

Let $w_{y,b} := \sum_{t\geq 1} \gamma^t E_{a_{1:t-1}, x_{1,t}} \left[\left(\prod_{i=1}^{t-1} c_i \right) \left(\pi(b|x_t) - \mu(b|x_t) c_t\left(b, \mathcal{F}_t\right) \right) I\left\{x_t = y\right\} \right] \geq 0$ by the hypothesis that $c \leq \frac{\pi}{\mu}$.
We have $\mathcal{R}Q(x,a) - Q^\pi(x,a) = \sum_{y,b} w_{y,b} \Delta Q(y,b)$.

Let's show that $\sum_{y,b} w_{y,b} \leq \gamma$.

$$\begin{aligned}
\sum_{y,b} w_{y,b} &= \sum_{t\geq 1} \gamma^t E_{a_{1:t-1},x_{1,t}} \left[\left(\prod_{i=1}^{t-1} c_i \right) \left(\sum_b \pi(b|x_t) - \mu(b|x_t) c_t(b, \mathcal{F}_t) \right) \right] \\
&= \sum_{t\geq 1} \gamma^t E_{a_{1:t-1},x_{1,t}} \left[\left(\prod_{i=1}^{t-1} c_i \right) \left(E_{a_t} \left[1 - c_t \left(a_t, \mathcal{F}_t \right) | \mathcal{F}_t \right] \right) \right] \\
&= \sum_{t\geq 1} \gamma^t E_{a_{1:t},x_{1,t}} \left[\left(\prod_{i=1}^{t-1} c_i \right) \left[1 - c_t \right] \right] \\
&= \gamma C - (C-1)
\end{aligned}$$

with $C := E_\mu \left[\sum_{t\geq 0} \gamma^t \left(\prod_{i=1}^{t} c_i \right) \right] \geq 1$. We have $\gamma C - (C-1) = C(\gamma - 1) + 1 \leq \gamma$ since $\gamma \leq 1$ and $C \geq 1$.
As a consequence,

$$\begin{aligned}
\mathcal{R}Q(x,a) - Q^\pi(x,a) &= \sum_{y,b} w_{y,b} \Delta Q(y,b) \\
&\implies \max_{x,a} |\mathcal{R}Q(x,a) - Q^\pi(x,a)| \leq \gamma \max_{y,b} |\Delta Q(x,a)| \\
&\implies \|RQ - Q^\pi\| \leq \gamma \|\Delta Q\|
\end{aligned}$$

which is **Theorem 1**.

3.2 Control

Definition. We say that a sequence of policies $(\pi_k,\ k \in N)$ is increasingly greedy with respect to a sequence $(Q_k,\ k \in N)$ if $\forall k,\ P^{\pi_{k+1}} Q^{k+1} \geq P^{\pi_k} Q^{k+1}$.

THe algorithms presented here need to converge to assume that the policies are increasingly greedy. By showing that two kinds of usual policies are increasingly greedy, we show that the assumption is achievable.

Lemma 2. Let (ϵ_k) be a non increasing sequence. Then the sequence of policies (π_k) which are ϵ_k-greedy w.r.t the sequence (Q_k) is increasingly greedy w.r.t that sequence.

Proof. We have :

$$\begin{aligned}
P^{\pi_{k+1}}Q_{k+1}(x,a) &= \sum_y p(y|x,a)\left((1-\varepsilon_{k+1})a_{greedy} + \epsilon_{k+1}a_{random}\right) \\
&= \sum_y p(y|x,a)\left[(1-\varepsilon_{k+1})\max_b Q_{k+1}(y,b) + \varepsilon_{k+1}\frac{1}{A}\sum_b Q_{k+1}(y,b)\right] \\
&\text{where } A = |\mathcal{A}| \\
&= \sum_y p(y|x,a)\left[\max_b Q_{k+1}(y,b) - \varepsilon_{k+1}\left(\max_b Q_{k+1}(y,b) - \frac{1}{A}\sum_b Q_{k+1}(y,b)\right)\right] \\
&\text{and } \max_b Q_{k+1}(y,b) - \frac{1}{A}\sum_b Q_{k+1}(y,b) \geq 0 \text{ and } \varepsilon_{k+1} \leq \varepsilon_k \\
&\geq \sum_y p(y|x,a)\left[(1-\varepsilon_k)\max_b Q_{k+1}(y,b) + \varepsilon_k\frac{1}{A}\sum_b Q_{k+1}(y,b)\right] \\
&= P^{\pi_k}Q_{k+1}
\end{aligned}$$

which is **Lemma 2**.

Lemma 3. Let (β_k) be a non-decreasing sequence of soft-max parameters. Then the sequence of policies (π_k) which are soft-max (with parameter β_k w.r.t. the sequence of functions (Q_k) is increasingly greedy w.r.t that sequence.

Proof. We define

$$\pi_\beta(b) = \frac{e^{\beta Q(y,b)}}{\sum_{b'} e^{\beta Q(y,b')}}$$

and

$$f(\beta) = \sum_b \pi_\beta(b)Q(y,b)$$

We show that f is increasing. f is C^∞ and

$$f'(\beta) = \sum_b \left(Q(y,b)\pi_\beta(b) - \pi_\beta(b)\sum_{b'} Q(y,b')e^{\beta Q(y,b)}\right)Q(y,b) = \mathrm{Var}_{b\sim\pi}\, Q(y,b) \geq 0$$

so f is non decreasing so $f(\beta_{k+1}) \geq f(\beta_k)$ and

$$\begin{aligned}
P^{\pi_{k+1}}Q_{k+1}(x,a) &= \sum_{y\in\mathcal{X}} P(y|x,a)\sum_b \pi_{\beta_{k+1}}(b)Q_{k+1}(y,b) \\
&\geq \sum_{y\in\mathcal{X}} P(y|x,a)\sum_b \pi_{\beta_k}(b)Q_{k+1}(y,b) \\
&= P^{\pi_{k+1}}Q_{k+1}(x,a)
\end{aligned}$$

which is **Lemma 3**.

Theorem 2. Consider an arbitrary sequence of behaviour policy (μ_k) (which may depend on (Q_k)) and a sequence of target policies which are increasingly greedy w.r.t the sequence (Q_k) defined by

$$Q_{k+1} = \mathcal{R}Q_k$$

with $c_s = c(a_s, x_s) \in \left[0, \frac{\pi_k(a_s|x_s)}{\mu_k(a_s|x_s)}\right]$ a Markovian. Assume the target policies π_k are ε_k -away from the greedy policies w.r.t. Q_k, in the sense that $\mathcal{T}^{\pi_k}Q_k \geq \mathcal{T}Q_k - \varepsilon_k \|Q_k\| e$ where e is the vector with ones. If we suppose that $\mathcal{T}^{\pi_0}Q_0 \geq \mathcal{T}Q_0$ then

$$\|Q_{k+1} - Q^*\| \leq \gamma \|Q_k - Q^*\| + \varepsilon_k \|Q_k\|$$

In consequence, if $\varepsilon_k \to 0$ then $Q_k \to Q^*$.

Proof. In the following, we will use

$$\left(P^{c\mu}Q\right)(x,a) := \sum_{x'}\sum_{a'} p\left(x'|x,a\right)\mu\left(a'|x'\right)c\left(a',x'\right)Q\left(x',a'\right)$$

Using $P^{c\mu}$ and the fact that we assumed that c_s was a Markovian, we write :

$$E_\mu\left(\prod_{s=1}^{t} c(a_s,x_s)\left[r_t + \gamma E_\pi(Q(x_{t+1},\cdot)) - Q(x_t,a_t)\right]\right) =$$

$$E_{a_{1:t-1},x_{1:t-1}}\left(\prod_{s=1}^{t-1} c(a_s,x_s) E_{a_t,x_t,x_{t+1}\sim P(\cdot|x_t,a_t)}\left(c(a_t,x_t)\left[r_t + \gamma E_{\pi_k}(Q(x_{t+1},\cdot)) - Q(x_t,a_t)\right]\right)\right) =$$

$$E_{a_{1:t-1},x_{1:t-1}}\left(\prod_{s=1}^{t-1} c(a_s,x_s)\sum_{x',a'} c(a',x')\mu(a'|x')P(x'|x_t,a_t)\left(\mathcal{T}^{\pi_k}Q - Q\right)(x_t,a_t)\right) =$$

$$E_{a_{1:t-1},x_{1:t-1}}\left(\prod_{s=1}^{t-1} c(a_s,x_s) P^{c\mu_k}\left(\mathcal{T}^{\pi_k}Q_k - Q_k\right)(x_t,a_t)\right)$$

By induction, we get

$$\mathcal{R}_k Q = Q + \sum_{t\geq 0}\gamma^t \left(P^{c\mu_k}\right)^t \left(\mathcal{T}^{\pi_k}Q - Q\right)$$

We notice that (since $\gamma\|P^{c\mu_k}\| < 1$

$$\left(\sum_{t\geq 0}\gamma^t P^{c\mu_k}\right)\left(I - \gamma P^{c\mu_k}\right) = I$$

So that

$$\mathcal{R}_k Q = Q + \left(I - \gamma P^{c\mu_k}\right)^{-1}\left(\mathcal{T}^{\pi_k}Q - Q\right)$$

The idea of the proof is to have an upper-bound and a lower-bound on $Q_k - Q^*$ which would gives us an upper-bound on $\|Q_k - Q^*\|$.

Upper bound on $Q_{k+1} - Q^*$

$$\begin{aligned}
Q_{k+1} - Q^* &= Q_k - Q^* + (I - \gamma P^{c\mu_k})^{-1} [\mathcal{T}^{\pi_k} Q_k - Q_k] \\
&= (I - \gamma P^{c\mu_k})^{-1} [\mathcal{T}^{\pi_k} Q_k - \mathcal{T} Q^* - \gamma P^{c\mu_k} (Q_k - Q^*)] \\
&\quad \text{where we used that } \mathcal{T} Q^* = Q^* \\
&\leq (I - \gamma P^{c\mu_k})^{-1} [\gamma P^{\pi_k} (Q_k - Q^*) - \gamma P^{c\mu_k} (Q_k - Q^*)] \\
&\quad \text{where we used that, by definition, } \mathcal{T} Q \geq r + \gamma P^{\pi} Q^* \ \forall \pi \text{ and } \mathcal{T}^{\pi_k} Q_k = r + \gamma P^{\pi_k} Q_k \\
&= A_k (Q_k - Q^*)
\end{aligned}$$

where $A_k := \gamma (I - \gamma P^{c\mu_k})^{-1} [P^{\pi_k} - P^{c\mu_k}]$.
$\forall k$ the coefficients of A_k are non-negative :

$$(P^{\pi_k} - P^{c\mu_k}) e(x,a) = \sum \sum p(x'|x,a) [\pi_k (a'|x') - c(a',x') \mu_k (a'|x')] \geq 0$$

by hypothesis since $\pi \geq c\mu$

Hence, $A_k = \sum_{t \geq 0} \gamma^t P^{c\mu_k t} (P^{\pi_k} - P^{c\mu_k})$ has non-negative coefficients.

Furthermore,

$$\begin{aligned}
A_k e &= \gamma \sum_{t \geq 0} \gamma^t (P^{c\mu_k})^t [P^{\pi_k} - P^{c\mu_k}] e \\
&= \gamma \sum_{t \geq 0} \gamma^t (P^{c\mu_k})^t e - \sum_{t \geq 0} \gamma^{t+1} (P^{c\mu_k})^{t+1} e \\
&\quad \text{since } P^{\pi_k} e = \sum_{x'} \sum_{a'} p(x'|x,a) \pi_k(a'|x') = 1 \\
&= e - (1 - \gamma) \sum_{t \geq 0} \gamma^t (P^{c\mu_k})^t e \\
&\quad \text{and } \sum_{t \geq 0} \gamma^t (P^{c\mu_k})^t e = e + \sum_{t \geq 1} \gamma^t (P^{c\mu_k})^t \geq e \\
&\leq \gamma e
\end{aligned}$$

We then write

$$\begin{aligned}
(Q_{k+1} - Q^*)(x,a) &= A_k (Q_k - Q^*)(x,a) \\
&\leq \|Q_k - Q^*\| A_k e(x,a) \\
&\quad \text{by non-negativity of the coefficients of } A_k \\
&\leq \gamma \|Q_k - Q^*\| e(x,a)
\end{aligned}$$

which implies $Q_{k+1} - Q^* \leq \gamma \|(Q_k - Q^*)\|$

Lower bound on $Q_{k+1} - Q^*$.
By hypothesis and by definition of $\mathcal{T}$ we have $\mathcal{T}^{\pi_k} Q_k \geq \mathcal{T} Q_k - \varepsilon_k \|Q_k\| e \geq \mathcal{T}^{\pi^*} Q_k - \varepsilon_k \|Q_k\| e$.
We can rewrite

$$\begin{aligned} Q_{k+1} - Q^* &= Q_{k+1} - \mathcal{T}^{\pi_k} Q_k + \mathcal{T}^{\pi_k} Q_k - \mathcal{T}^{\pi^*} Q_k + \mathcal{T}^{\pi^*} Q_k - \mathcal{T}^{\pi^*} Q^* \\ &\quad \text{where we used that } Q^* \text{ is a fixed point of } \mathcal{T}^{\pi *} \\ &\geq Q_{k+1} - \mathcal{T}^{\pi_k} Q_k + \mathcal{T}^{\pi^*} Q_k - \varepsilon_k \|Q_k\| e - \mathcal{T}^{\pi^*} Q_k + \mathcal{T}^{\pi^*} Q_k - \mathcal{T}^{\pi^*} Q^* \\ &\geq Q_{k+1} - \mathcal{T}^{\pi_k} Q_k + \gamma P^{\pi^*} (Q_k - Q^*) - \varepsilon_k \|Q_k\| e \end{aligned}$$

by the previous inequalities.

and then, using that $Q_{k+1} = \mathcal{T}^{\pi_k} Q_k + \gamma P^{c\mu_k} (I - \gamma P^{c\mu_k})^{-1} (\mathcal{T}^{\pi_k} Q_k - Q_k)$, we get

$$Q_{k+1} - Q^* \geq \gamma P^{c\mu_k} (I - \gamma P^{c\mu_k})^{-1} (\mathcal{T}^{\pi_k} Q_k - Q_k) + \gamma P^{\pi^*} (Q_k - Q^*) - \varepsilon_k \|Q_k\|$$

We know want a lower bound on $\mathcal{T}^{\pi_k} Q_k - Q_k$ to have our lower-bound on $Q_{k+1} - Q^*$.

Lower bound on $\mathcal{T}^{\pi_k} Q_k - Q_k$. We have (because ε_k is increasingly greedy w.r.t Q_k) that :

$$\begin{aligned} \mathcal{T}^{\pi_{k+1}} Q_{k+1} - Q_{k+1} &\geq \mathcal{T}^{\pi_k} Q_{k+1} - Q_{k+1} \\ &= \mathcal{T}^{\pi_k} \mathcal{R}_k Q_k - \mathcal{R}_k Q_k \\ &= r + \gamma P^{\pi_k} \mathcal{R}_k Q_k - \mathcal{R}_k Q_k \\ &= r + (\gamma P^{\pi_k} - I) \mathcal{R}_k Q_k \\ &= r + (\gamma P^{\pi_k} - I) \left[Q_k + (I - \gamma P^{c\mu_k})^{-1} (\mathcal{T}^{\pi_k} Q_k - Q_k) \right] \\ &= \mathcal{T}^{\pi_k} Q_k - Q_k + (\gamma P^{\pi_k} - I) (I - \gamma P^{c\mu_k})^{-1} (\mathcal{T}^{\pi_k} Q_k - Q_k) \\ &= B_k (\mathcal{T}^{\pi_k} Q_k - Q_k) \end{aligned}$$

where $B_k := \gamma [P^{\pi_k} - P^{c\mu_k}] (I - \gamma P^{c\mu_k})^{-1}$.
As we have showed earlier, B_k only has non-negative elements (same proof as for A_k). With the hypothesis that $\mathcal{T}^{\pi_0} Q_0 - Q_0 \geq 0$ that makes that $\mathcal{T}^{\pi_k} Q_k - Q_k \geq 0$ $\forall k$ and finally

$$Q_{k+1} - Q^* \geq \gamma P^{\pi^*} (Q_k - Q^*) - \varepsilon_k \|Q_k\|$$

Using that $\|P^{\pi^*}(Q_k - Q^*)\| \leq \|(Q_k - Q^*)\|$ because we have seen that P^π is non-expansive $\forall \pi$ policy, we get that :

$$\|Q_{k+1} - Q^*\| \leq \gamma \|Q_k - Q^*\| + \varepsilon_k \|Q_k\|$$

If $\varepsilon_k \to 0$, we show that Q_k is bounded. For $k > k_0$ big enough, we have $\varepsilon_k < \frac{1-\gamma}{2}$. We then have :

$$\begin{aligned}\|Q_{k+1}\| &= \|Q_{k+1} - Q^* + Q^*\| \\ &\leq \|Q^*\| + \gamma \|Q_k - Q^*\| + \varepsilon_k \|Q_k\| \\ &\leq (1+\gamma)\|Q^*\| + \frac{1+\gamma}{2}\|Q_k\|\end{aligned}$$

which implies that $\|Q_k\|$ is bounded, by taking the limsup. It follows that $\limsup Q_k = Q^*$.

3.3 Online algorithm

We now want to show that the Retrace(λ) algorithm can learn from sample trajectories.

Definition. A sequence (α_k) obeys the Robbins-Monro conditions if :

1. $\sum_k \alpha_k = \infty$
2. $\sum_k \alpha_k^2 < \infty$

Notation : In the context of the Retrace(λ) algorithm we rewrite

$$P^{c\mu} = \lambda P^{\pi \wedge \mu} Q(x,a) := \sum_y \sum_b \min(\pi(b|y), \mu(b|y)) Q(y,b)$$

which allows us to write $\mathcal{R}Q = Q + (I - \lambda\gamma P^{\pi\wedge\mu})^{-1}(\mathcal{T}^\pi Q - Q)$ because $c(b,y)\mu(b|y)\pi(b|y) = \lambda \min(\mu(b|y)\pi(b|y))$.

Theorem 3. Consider a sequence of sample trajectories with the k^{th} trajectory $x_0, a_0, r_0, x_1, a_1, r_1 ...$ generated by following $\mu_k : a_t \sim \mu(\cdot, x_t)$. For each, (x,a), with s being the time of fist occurrence of (x,a), update

$$Q_{k+1}(x,a) \leftarrow Q_k(x,a) + \alpha_k \sum_{t \geq s} \delta_t^{\pi_k} \sum_{j=s}^{t} \gamma^{t-j} \left(\prod_{i=j+1}^{t} c_i \right) I\{x_j = x, a_j = a\}$$

where $\delta_t^{\pi_k} := r_t + \gamma E_{\pi_k} Q_k(x_{t+1}, \cdot) - Q_k(x_t, a_t), \alpha_k = \alpha_k(x_s, a_s)$. We consider the Retrace(λ) algorithm where $c_i = \lambda \min\left(1, \frac{\pi(a_i|x_i)}{\mu(a_i|x_i)}\right)$. Assume that (π_k) are increasingly greedy w.r.t (Q_k) and are ε_k-away from the greedy policies π_{Q_k} i.e $\max_x \|\pi_k(\cdot, x) - \pi_{Q_k}(\cdot, x)\|_1 \leq \varepsilon_k$ with $\varepsilon_k \to 0$. Assume that P^{π_k} and $P^{\pi_k \wedge \mu_k}$ asymptotically commute i.e $\lim_k \|P^{\pi_k} P^{\pi_k \wedge \mu_k} - P^{\pi_k \wedge \mu_k} P^{\pi_k}\| = 0$. Assume further that :

1. all states and actions are visited infinitely often: $\sum_{t \geq 0} P\{x_t, a_t = x, a\} \geq D > 0$

2. the sample trajectories are finite in terms of the second moment of their lengths T_k : $E_{\mu_k} T_k^2 < \infty$

3. the stepsizes obey the usual Robbins-Munro conditions.

then $Q_k \to Q^*$ a.s.

To prove that theorem, we need a more general theorem :

Theorem 4. Consider the algorithm $Q_{k+1}(x,a) = (1-\alpha_k(x,a))\, Q_k(x,a) + \alpha_k(x,a)\,(\mathcal{R}_k Q_k(x,a) + \omega_k(x,a) + v_k(x,a))$ and assume that

1. ω_k is a centered, $\mathcal{F}_k$ -measurable noise term of bounded variance

2. v_k is bounded from above by $\theta_k\,(\|Q_k\|+1)$

then under the same assumptions as **Theorem 3**, $Q_k \to Q^*$ a.s.

Proof. The proof is very similar to the one of **Theorem 2**. We denote by $\mathcal{R}$ the $\mathcal{R}_k$ operator.

Upper bound on $\mathcal{R}Q_k - Q^*$. By applying exactly the same process as in the upper-bound in **Theorem 2**, we get

$$\begin{aligned}\mathcal{R}Q_k - Q^* &\leq \lambda\gamma\,\|Q_k - Q^*\|\,e \\ &\leq \gamma\,\|Q_k - Q^*\|\,e\end{aligned}$$

Lower bound on $\mathcal{R}Q_k - Q^*$. Taking $\lambda P^{\pi\wedge\mu}$ instead of $P^{c\mu}$, and applying the same result as in **Theorem 2**, we get

$$\begin{aligned}\mathcal{R}Q_k - Q^* \geq& \gamma\lambda P^{\pi_k\wedge\mu_k}\left(I - \gamma\lambda P^{\pi_k\wedge\mu_k}\right)^{-1}\left(\mathcal{T}^{\pi_k}Q_k - Q_k\right) \\ &+ \gamma P^{\pi^*}\left(Q_k - Q^*\right) - \varepsilon_k\,\|Q_k\|\end{aligned}$$

Lower-bound on $\mathcal{T}^{\pi_k}Q_k - Q_k$. π_k being increasingly greedy w.r.t (Q_k), we write :

$$\begin{aligned}\mathcal{T}^{\pi_{k+1}}Q_{k+1} - Q_{k+1} &\geq \mathcal{T}^{\pi_k}Q_{k+1} - Q_{k+1} \\ &= (1-\alpha_k)\,\mathcal{T}^{\pi_k}Q_k + \alpha_k\mathcal{T}^{\pi_k}\left(\mathcal{R}Q_k + \omega_k + v_k\right) - Q_{k+1} \\ &\quad \text{where we used the definition of } Q_{k+1} \text{ in the online algorithm.} \\ &= (1-\alpha_k)\left(\mathcal{T}^{\pi_k}Q_k - Q_k\right) + \alpha_k\left[\mathcal{T}^{\pi_k}\mathcal{R}Q_k - \mathcal{R}Q_k + \omega_k' + v_k'\right]\end{aligned}$$

with $\omega_k' = (\gamma P^{\pi_k} - I)\omega_k$ and $v_k' := (\gamma P^{\pi_k} - I)\,v_k$. ω_k' is still centered (we just apply a linear transformation on ω_k) and its variance is still bounded by $\|I - \gamma P^{\pi_k}\|^2\,\mathrm{Var}(\omega_k)$. We also have that $\|v_k'\| \leq 2\|v_k\|$. As a consequence, ω'

and v' still verify the same assumptions as ω and v.
Using the definition of $\mathcal{T}^{\pi_k}$:

$$\begin{aligned}
\mathcal{T}^{\pi_k}\mathcal{R}Q_k - \mathcal{R}Q_k &= r + (\gamma P^{\pi_k} - I)\,\mathcal{R}Q_k \\
&= r + (\gamma P^{\pi_k} - I)\left[Q_k + (I - \gamma\lambda P^{\pi_k \wedge \mu_k})^{-1}\,(\mathcal{T}^{\pi_k}Q_k - Q_k)\right] \\
&= \mathcal{T}^{\pi_k}Q_k - Q_k + (\gamma P^{\pi_k} - I)\,(I - \gamma\lambda P^{\pi_k \wedge \mu_k})^{-1}\,(\mathcal{T}^{\pi_k}Q_k - Q_k) \\
&= \left[I + (\gamma P^{\pi_k} - I)\,(I - \gamma\lambda P^{\pi_k \wedge \mu_k})^{-1}\right](T^{\pi_k}Q_k - Q_k) \\
&= \gamma\,(P^{\pi_k} - \lambda P^{\pi_k \wedge \mu_k})\,(I - \gamma\lambda P^{\pi_k \wedge \mu_k})^{-1}\,(T^{\pi_k}Q_k - Q_k) \\
&= B_k\,(T^{\pi_k}Q_k - Q_k)
\end{aligned}$$

with $B_k := \gamma\,(P^{\pi_k} - \lambda P^{\pi_k \wedge \mu_k})\,(I - \gamma\lambda P^{\pi_k \wedge \mu_k})^{-1}$.
As a consequence, with $\xi_k := \mathcal{T}^{\pi_k}Q_k - Q_k$, we have :

$$\xi_{k+1} \geq (1 - \alpha_k)\xi_k + \alpha_k(B_k\xi_k + \omega'_k + v'_k)$$

We are going to use the assumption of asymptotical commutativity to show that A_k and B_k are quite close.
First we notice that, $\forall t$,

$$\begin{aligned}
\sum_{s=0}^{t-1} V^s(UV - VU)V^{t-s-1} + V^tU &= \sum_{s=0}^{t-1} V^sUV^{t-s} - V^{s+1}UV^{t-(s+1)} + V^tU \\
&\qquad \text{which is telescopical} \\
&= UV^t - V^tU + V^tU \\
&= UV^t
\end{aligned}$$

Any transition or sub-transition matrix verifies $\|U\| \leq 1$. That allows us to write (with $\lambda\gamma < 1$):

$$\begin{aligned}
U(I - \lambda\gamma V)^{-1} &= \sum_{t\geq 0}(\lambda\gamma)^tUV^t \\
&= \sum_{t\geq 0}(\lambda\gamma)^t\left[\sum_{s=0}^{t-1} V^s(UV - VU)V^{t-s-1} + V^tU\right] \\
&= (I - \lambda\gamma V)^{-1}U + \sum_{t\geq 0}(\lambda\gamma)^t\sum_{s=0}^{t-1} V^s(UV - VU)V^{t-s-1} \\
&\text{where we wrote } \sum_{t\geq 0}(\gamma\lambda)^tV^tU = (I - \lambda\gamma V)^{-1}U
\end{aligned}$$

We then have :

$$\begin{aligned}\|A_k - B_k\| &\leq \gamma \sum_{t\geq 0} (\lambda\gamma)^t \sum_{s=0}^{t-1} \eta_k \|\boldsymbol{P}^{\pi_k \wedge \mu_k}\|^{t-1} \\ &= \gamma \sum_{t\geq 0} (\lambda\gamma)^t t\eta_k \\ &\leq \gamma \frac{1}{(1-\lambda\gamma)^2} \eta_k\end{aligned}$$

Then we have :

$$\begin{aligned}\xi_{k+1} &\geq (1-\alpha_k)\xi_k + \alpha_k \left(A_k\xi_k + (B_k - A_k)\,\xi_k + \omega'_k + v'_k\right) \\ &\geq (1-\alpha_k)\xi_k + \alpha_k \left(A_k\xi_k + +\omega'_k + v''_k\right) \\ &\text{where } v''_k := v'_k + \gamma \sum_{t\geq 0} t(\lambda\gamma)^t \eta_k \,\|\xi_k\| e\end{aligned}$$

Since $\|\xi_k\|$ is bounded by a constant times $\|Q_k\|$ and $\eta_k \to 0$, v'' verify the same assumptions as v' and v.
Let's define

$$\xi'_{k+1} = (1-\alpha_k)\,\xi'_k + \alpha_k \left(A_k\xi'_k + \omega'_k + v''_k\right)$$

. We have that :

1. The matrice A_k are non-negative.
2. A_k is a γ-contraction.
3. The noise ω'_k is $\mathcal{F}_k$ measurable, centered and verifies the bounded variance assumption.
4. v''_k is bounded above by $(1+\gamma)\theta'_k \left(\|Q_k\| + 1\right)$ with $\theta'_k \to 0$.
5. The Robbins-Monro conditions are verified

Using Bertsekas & Tsitsiklis *Proposition 4.5* from *Neuro-dynamic programming, (1996)*, we get that $\xi'_k \to 0$. By induction it is clear that $\forall k, \xi'_k \leq \xi_k$ which gives $\liminf_k \xi_k \geq 0$. We deduce that :

$$\begin{aligned}\|\mathcal{R}Q_k - Q^*\| \leq \max(&\|\gamma\lambda P^{\pi_k\wedge\mu_k} \left(I - \gamma\lambda P^{\pi_k\wedge\mu_k}\right)^{-1} \left(\mathcal{T}^{\pi_k}Q_k - Q_k\right) + && (3)\\ &\gamma P^{\pi^*}\left(Q_k - Q^*\right) - \varepsilon_k \|Q_k\|\,\|, \|\gamma \|Q_k - Q^*\|\, e\|) && (4)\\ &\leq \gamma \|Q_k - Q^*\| + O\left(\varepsilon_k \|Q_k\|\right) + O\left(\xi_k\right) && (5)\\ &\text{where we used that } \sum_{t\geq 0} (\gamma\lambda)^t (P^{\pi_k\wedge\mu_k})^t = (I - \gamma\lambda P^{\pi_k\wedge\mu_k})^{-1} && (6)\\ &\text{which implies that } (I - \gamma\lambda P^{\pi_k\wedge\mu_k})^{-1} \text{ is bounded.} && (7)\end{aligned}$$

The online algorithm is defined by :

$$Q_{k+1} = (1-\alpha_k)Q_k + \alpha_k \left(\mathcal{R}_k Q_k + \omega_k + v_k\right)$$

If we call $\omega'_k = O(\varepsilon_k \|Q_k\|)$ and $v'_k = v_k + O(\xi_k)$, we admit that *Proposition 4.5* from *Neuro-Dynamic programming* by *Bertsekas & Tsitsiklis* allows us to conclude that $Q_k \to Q^*$ a.s..

Proof of Theorem 3. We are going to use **Theorem 4**. For that, we need to write

$$Q_{k+1}(x,a) \leftarrow Q_k(x,a) + \alpha_k \sum_{t \geq s} \delta_t^{\pi_k} \sum_{j=s}^{t} \gamma^{t-j} \left(\prod_{i=j+1}^{t} c_i \right) I\{x_j, a_j = x, a\}$$

under the form

$$Q_{k+1}(x,a) = (1 - \alpha_k(x,a))\, Q_k(x,a) + \alpha_k(x,a) \left(\mathcal{R}_k Q_k(x,a) + \omega_k(x,a) + v_k(x,a)\right)$$

By over-bounding $\gamma^{t-j} \left(\prod_{i=j+1}^{t} c_i \right) I\{(x_j,a_j) = (x_s,a_s)\}$ by 1 and using the assumption on the second order moment of T_k, we get that

$$E\left[\sum_{t \geq s} z_{s,t}^k | \mathcal{F}_k \right] < E\left[T_k^2 | \mathcal{F}_k \right] < \infty$$

Moreover :

$$E_{\mu_k}\left[\sum_{t \geq s} \delta_t^{\pi_k} z_{s,t}^k \right] =$$

$$E_{\mu_k}\left[\sum_{t \geq s} \left(r_t + \gamma E_{\pi_k} Q_k^o(x_{t+1}, \cdot) - Q_k^o(x_t, a_t) \right) \sum_{j=s}^{t} \gamma^{t-j} \left(\prod_{i=j+1}^{t} c_i \right) I\{(x_j,a_j) = (x_s,a_s)\} \right] =$$

$$D_k(x_s,a_s) \left(\mathcal{R}_k Q_k(x_s,a_s) - Q(x_s,a_s) \right)$$

where we used the expression of $\mathcal{R}_k$ and where $D_k(x_s,a_s) := \sum_{t \geq s} P\{(x_t,a_t) = (x_s,a_s)\} > 0$ by hypothesis.
To get the desired form, we call $\Delta_s := \sum_{t \geq s} \delta_t^{\pi_k} z_{s,t}^k$ and we write :

$$\begin{aligned}
Q_{k+1}^o &= Q_k^o + \alpha_k \Delta_s \\
&= Q_k^o + \alpha_k D_k D_k^{-1} (\Delta_s - E_{\mu_k}\Delta_s + E_{\mu_k}\Delta_s) \\
&= Q_k^o + \tilde{\alpha_k} D_k^{-1} (\Delta_s - E_{\mu_k}(\Delta_s)) + \tilde{\alpha_k} (\mathcal{R}_k Q_k^o - Q_k) \\
&= (1 - \tilde{\alpha_k}) Q_k^o + \tilde{\alpha_k} (\mathcal{R}_k Q_k^o + w_k + v_k)
\end{aligned}$$

where :

$$Q^o_{k+1}(x_s,a_s) \leftarrow (1-\tilde{\alpha}_k) Q^o_k(x_s,a_s) + \tilde{\alpha}_k \left(\mathcal{R}_k Q^o_k(x_s,a_s) + \omega_k(x_s,a_s) + v_k(x_s,a_s)\right)$$

$$\omega_k(x_s,a_s) := (D_k)^{-1} \left(\sum_{t\geq s} \delta_t^{\pi_k} z^k_{s,t} - E_{\mu_k}\left[\sum_{t\geq s} \delta_t^{\pi_k} z^k_{s,t} \right] \right)$$

$$v_k(x_s,a_s) := (\tilde{\alpha}_k)^{-1} \left(Q^o_{k+1}(x_s,a_s) - Q_{k+1}(x_s,a_s) \right)$$

$$\tilde{\alpha}_k := \alpha_k D_k$$

Conditions on $\tilde{\alpha_k}$:It is clear, since $D_k \geq D > 0$ that $\tilde{\alpha_k}$ that $\tilde{\alpha_k}$ still fulfills Robbins & Monro conditions.
Conditions on v : We admit that v follows the conditions of **Theorem 4**, which is a consequence of *Proposition 5.2* of *Bersketas & Tsitsiklis*
Conditions on w : We use the fact that r is bounded,that the trajectory are finite and that $ET_k^2 < \infty$.

That allows us to apply **Theorem 4** to conclude that $Q^o_k \to Q^*$ as $k \to \infty$ with probability 1.